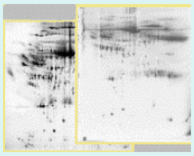
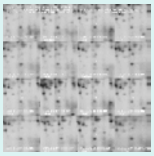
Overview:



Open2Dprot

The Open N-Dimensional Proteomics Project



http://open2dprot.sourceforge.net/

Revised: 09-12-2004, P. Lemkin

Overview

- What is the Open2Dprot project?
- What is open source?
- Why are we using it for this project?
- Project goals
- Open source resources
- Development plan
  - initial and second phases
  - community standard proteomics DB schemas
  - technology design
- Bioinformatics community core-support

The Open2Dprot Project

Open2Dprot is an open-source project for the development of n-dimensional proteomics exploratory data analysis bioinformatic tools.

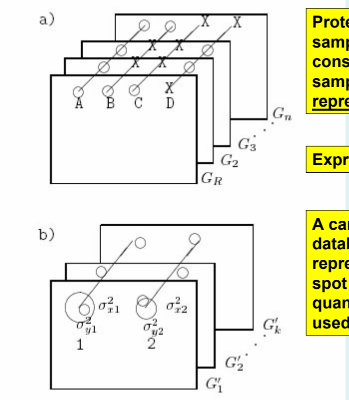
The tools can be used for analyzing quantified protein expression data across multiple n-D samples from research experiments.

The tools could be adapted for use with a variety of quantified 2-D or n-dimensional protein separation sources of expression data.

Proteomic Separation Methods

- 2D-PAGE (P. O'Farrell, 1975) pIe vs Mm (mass), 2D-gels
  2D LC-MS retention-times vs m/z (mass)
  2D IPG-MS pIe vs m/z (mass)
  n-D (e.g., LC-MS*MS*MS …)
- All share a common paradigm: proteins separated by orthogonal features
- Some methods are semi-quantitative
- Data represented as protein expression profiles lends itself to exploratory data analysis
- Open2Dprot could be used as basis for a broader set of integrated tools

Composite Samples Database (CSD) Paradigm



Proteomic composite samples database (CSD) consisting of a set of n samples G_1, G_2, …,G_n with representative sample $G_{r=}G_1$

Expression profiles A,B,C, ...

A canonical sample database is a statistical representation of the CSD spot geometry and quantification that could be used for data mining

in Lemkin *et al.*, *Computers Biomedical Research*, 1981

Why 2D-Gels Now?

- 2D-PAGE was not widely used until recently due to:
  - limitations in identifying spots differentially expressed
  - difficulty resolving and detecting specialized classes of proteins (e.g., basic proteins, membrane proteins, low abundance proteins)
- Today, 2D-PAGE is often used as prescreening stage for mass-spectrometry to identify spots found in differential analysis
- Improved resolution: zoom 2D-gels, new pre-fractionation methods
- There are other protein separation techniques that could use these 2D-gel and recent DNA-microarray database analysis paradigms including 2D LC-MS

Why Open Source?

“The basic idea behind open source is very simple: When programmers can read, redistribute, and modify the source code for a piece of software, the software evolves. People improve it, people adapt it, people fix bugs. And this can happen at a speed that, if one is used to the slow pace of conventional software development, seems astonishing.**”**

“We in the open source community have learned that this rapid evolutionary process produces better software than the traditional closed model, in which only a very few programmers can see the source and everybody else must blindly use an opaque block of bits.**”**

From the Open Source Initiative (OSI)

http://www.opensource.org/

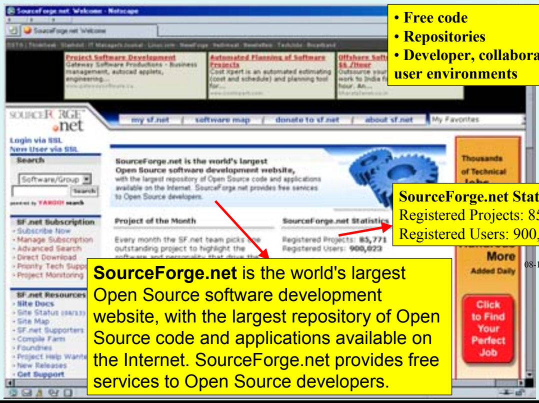
Why an Open-Source nD-Data Proteomics Effort?

- “***An open-source project can be advantageous to the community at large, since there is a far greater likelihood of progress in algorithm design in an academic style collaboration than a closed-source business model***.”
- Researchers can more rapidly adapt new methods to existing software without waiting for release of commercial products
- Use contributed expertise and code of proteomics experts and bioinformaticians to help build and test open software
- Algorithms more transparent, so researchers can verify results more easily
- More opportunity to share data in standard non-proprietary

Why Open Source Proteomics? (continued)

- No expensive software licenses required - reduces deployment costs within large organizations and small labs
- Using proper open-source licenses can encourage adoption and collaboration by commercial interests
- Many free open-source repositories available
- Repositories offer tools to support collaboration, software development and distribution

Open Source Repositories - E.g., SourceForge.Net



SourceForge.net is the world's largest Open Source software development website, with the largest repository of Open Source code and applications available on the Internet. SourceForge.net provides free services to Open Source developers.

Project of the Month

Every month the SF.net team picks one outstanding project to highlight the software and personality that drive the

SourceForge.net Statistics

Registered Projects: 85,771
Registered Users: 900,023

- **Free code**
- **Repositories**
- **Developer, collaborator, user environments**

SourceForge.net Statistics
Registered Projects: 85,771
Registered Users: 900,023

08-15-2004

SourceForge.net is the world's largest Open Source software development website, with the largest repository of Open Source code and applications available on the Internet. SourceForge.net provides free services to Open Source developers.

Open2Dprot - Project Goals

- An international community effort to create an open-source n-D quantitative data analysis system
- A stand-alone downloadable system that can connect to DBs
- Could be used for data mining protein expression across sets of samples from researcher’s experiments to investigate and find significant protein expression from multiple experiments
- Will provide integrated set of software tools, analysis methods and data structures for quantitative and system biology protein expression
- Will handle protein expression data from 2D-gel, 2D LC-MS, and other protein separation methods

Using Open Source Resources

- Initially, hosted and developed on SourceForge.Net repository at **open2dprot.sourceforge.net**
- This Web site discusses the current Open2Dprot software development plan
- Use the same open-source development methodology used in our Java/R-based MAExplorer **maexplorer.sourceforge.net** DNA microarray data-mining software
- Open2Dprot could later reside as part of **HUPO.org** analysis Web site integrated with other tools relating to mass spectrometry, dye multiplexing, protein arrays, Internet proteomic databases, etc.

Development Plan

- Open2Dprot is being written in Java and R languages using XML and MySQL RDBMS - modern modular open-source technologies aiding portability and extensibility

- Initial phase: Open2Dprot is being derived from refactored code
  a) parts of NCI GELLAB-II system - the C-language / Unix / X-windows 1993 version (www.lecb.ncifcrf.gov/gellab),
  b) from other open source proteomics and bioinformatics projects
  c) Java / R / plugins from MAExplorer and R data-mining software

- Second phase: extended with other donated 2D-gel, LC-MS^N and other analysis and related proteomics software codes with additional efforts by the research community

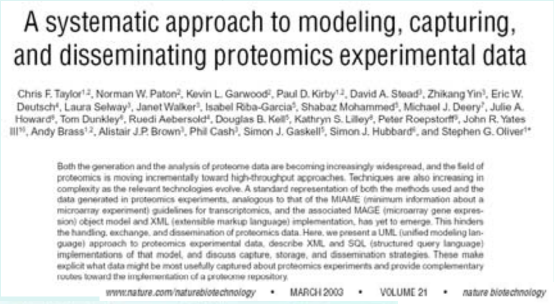
Development Plan (cont.)

- Work with proteomics standardization groups (MIAPE - formerly PEDRo, PSI, HUPO, and others) to develop and use a standard database schema

- Encourage research community to help expand, extend and integrate basic paradigm with other related protein separation methods and data analysis methods

- During initial phase, we especially welcome suggestions for modifying this agenda for Open2Dprot as well as core-bioinformatics developers offering to help with the project

'PEDRo' - Proteomic Experiment Data Repository Schema Standard

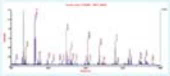


A systematic approach to modeling, capturing, and disseminating proteomics experimental data

Chris F. Taylor[1,2], Norman W. Paton[2], Kevin L. Garwood[2], Paul D. Kirby[1,2], David A. Stead[3], Zhikang Yin[3], Eric W. Deutsch[4], Laura Selway[3], Janet Walker[3], Isabel Riba-Garcia[5], Shabaz Mohammed[5], Michael J. Deery[7], Julie A. Howard[8], Tom Dunkley[4], Ruedi Aebersold[4], Douglas B. Kell[5], Kathryn S. Lilley[8], Peter Roepstorff[9], John R. Yates III[10], Andy Brass[1,2], Alistair J.P. Brown[3], Phil Cash[3], Simon J. Gaskell[5], Simon J. Hubbard[6], and Stephen G. Oliver[1]*

Both the generation and the analysis of proteome data are becoming increasingly widespread, and the field of proteomics is moving incrementally toward high-throughput approaches. Techniques are also increasing in complexity as the relevant technologies evolve. A standard representation of both the methods used and the data generated in proteomics experiments, analogous to that of the MIAME (minimum information about a microarray experiment) guidelines for transcriptomics, and the associated MAGE (microarray gene expression) object model and XML (extensible markup language) implementation, has yet to emerge. This hinders the handling, exchange, and dissemination of proteomics data. Here, we present a UML (unified modeling language) approach to proteomics experimental data, describe XML and SQL (structured query language) implementations of that model, and discuss capture, storage, and dissemination strategies. These make explicit what data might be most usefully captured about proteomics experiments and provide complementary routes toward the implementation of a proteome repository.

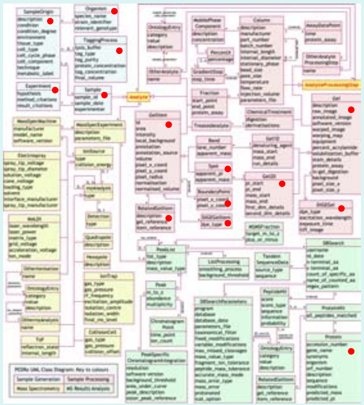
www.nature.com/naturebiotechnology • MARCH 2003 • VOLUME 21 • nature biotechnology



psidev.sourceforge.net

pedro.sourceforge.net

MIAPE (PEDRo) UML Schema n-D Data Classes



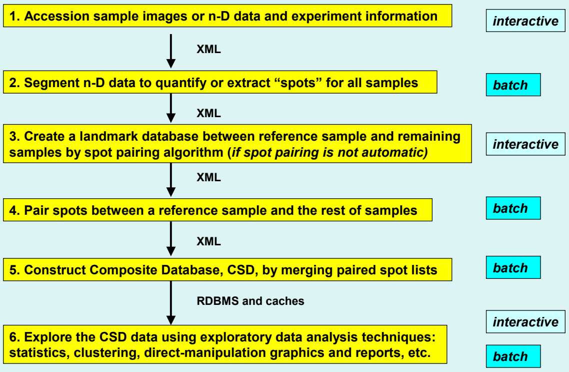
- Classes that could be used with 2D-gels

  Additional fields / classes are needed for Open2Dprot

  in Taylor et.al., *Nature Biotechnology*, March 2003.

PEDRo has been renamed MIAPE "Minimal Information About a Proteomics Experiment" (Oct. 2003, HUPO-II) by EMBL-EBI

Basic Open n-D Analysis Pipeline



1. Accession sample images or n-D data and experiment information — *interactive*

XML

2. Segment n-D data to quantify or extract "spots" for all samples — *batch*

XML

3. Create a landmark database between reference sample and remaining samples by spot pairing algorithm (*if spot pairing is not automatic*) — *interactive*

XML

4. Pair spots between a reference sample and the rest of samples — *batch*

XML

5. Construct Composite Database, CSD, by merging paired spot lists — *batch*

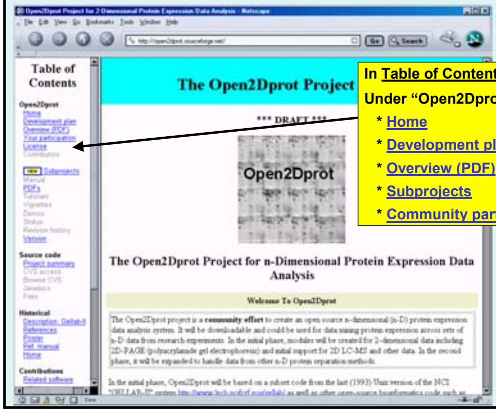
RDBMS and caches

6. Explore the CSD data using exploratory data analysis techniques: statistics, clustering, direct-manipulation graphics and reports, etc. — *interactive* *batch*

Initial Open n-D Data-Mining Tools

- **Accession n-D sample images or n-D data and experiment data**
- **Quantify 'spots' from sample images or peptide clusters**
- ***Pair spots between samples and a reference sample***
- **Construct composite sample database for exploratory data analysis**
- **Manage subsets of proteins in the database**
- **Manage replicate samples and condition sets of samples**
- **Analyze expression profiles for multiple conditions**
- **Data-filter protein sets by statistics, clustering, set membership**
- **Direct-manipulation of data in graphics, spreadsheets**
- **Integrate R language statistical, clustering, classifiers, class prediction, and other methods**
- **Integrate access to Internet proteomic/genomic/function data servers for user-specified protein sets**

Home: http://open2dprot.sourceforge.net/



Open2Dprot Project for 2-Dimensional Protein Expression Data Analysis - Netscape

http://open2dprot.sourceforge.net/

Table of Contents

Open2Dprot
Home
Development plan
Overview (PDF)
Your participation
License
Contributors

Subprojects
Manual
PDFs
Tutorials
Vignettes
Demos
Status
Revision history
Version

Source code
Project summary
CVS access
Browse CVS
Javadocs
Files

Historical
Description: Gellab-II
References
Poster
Ref. manual
Home

Contributions
Related software

The Open2Dprot Project

*** DRAFT ***

Open2Dprot

The Open2Dprot Project for n-Dimensional Protein Expression Data Analysis

Welcome To Open2Dprot

The Open2Dprot project is a **community effort** to create an open source n-dimensional (n-D) protein expression data analysis system. It will be downloadable and could be used for data mining protein expression across sets of n-D data from research experiments. In the initial phase, modules will be created for 2-dimensional data including 2D-PAGE (polyacrylamide gel electrophoresis) and initial support for 2D LC-MS and other data. In the second phase, it will be expanded to handle data from other n-D protein separation methods.

In the initial phase, Open2Dprot will be based on a subset code from the last (1993) Unix version of the NCI "GELLAB-II" system http://www.lecb.ncifcrf.gov/gellab/ as well as other open-source bioinformatics code such as

In Table of Contents, see:

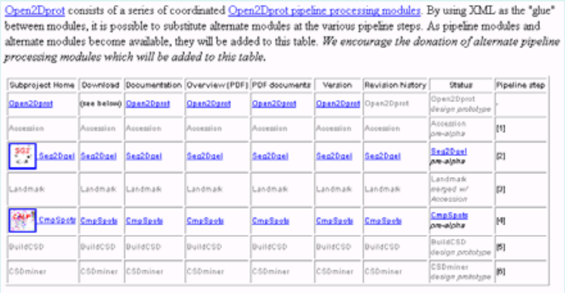
Under "Open2Dprot"

- * Home
- * Development plan
- * Overview (PDF)
- * Subprojects
- * Community participation

Bioinformatics Community Core-Support

1. Initial phase: bioinformatics core-developers to help refactor code to modular (Java / R / XML / MySQL-RDBMS) paradigm
2. A few senior bioinformatics core-developers to take on managerial and design roles (a long-term goal is to have multiple "project managers" in various proteomics specialties)
3. Active research groups to beta-test system with their data
4. Help with subsequent extension/integration with other protein separation methods software/databases, statistics, data mining, etc.
5. Contributions of alternative computation modules for analysis pipeline - e.g., spot quantification, pairing, statistical analysis, etc.

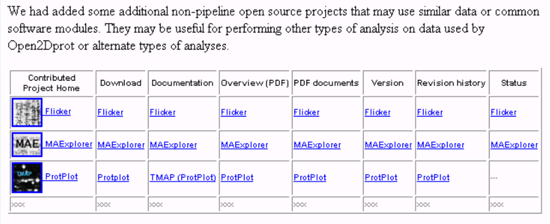
Open2Dprot Pipeline Subprojects



Open2Dprot consists of a series of coordinated Open2Dprot pipeline processing modules. By using XML as the "glue" between modules, it is possible to substitute alternate modules at the various pipeline steps. As pipeline modules and alternate modules become available, they will be added to this table. *We encourage the donation of alternate pipeline processing modules which will be added to this table.*

Subproject Home	Download	Documentation	Overview (PDF)	PDF documents	Version	Revision history	Status	Pipeline step
Open2Dprot	(see below)	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot	Open2Dprot *design prototype*	-
Accession	Accession	Accession	Accession	Accession	Accession	Accession	Accession *pre-alpha*	[1]
Seg2Dgel	Seg2Dgel	Seg2Dgel	Seg2Dgel	Seg2Dgel	Seg2Dgel	Seg2Dgel	Seg2Dgel *pre-alpha*	[2]
Landmark	Landmark	Landmark	Landmark	Landmark	Landmark	Landmark	Landmark *merged w/ Accession*	[3]
CmpSpots	CmpSpots	CmpSpots	CmpSpots	CmpSpots	CmpSpots	CmpSpots	CmpSpots *pre-alpha*	[4]
BuildCSD	BuildCSD	BuildCSD	BuildCSD	BuildCSD	BuildCSD	BuildCSD	BuildCSD *design prototype*	[5]
CSDminer	CSDminer	CSDminer	CSDminer	CSDminer	CSDminer	CSDminer	CSDminer *design prototype*	[6]

Associated or Related Projects



We had added some additional non-pipeline open source projects that may use similar data or common software modules. They may be useful for performing other types of analysis on data used by Open2Dprot or alternate types of analyses.

Contributed Project Home	Download	Documentation	Overview (PDF)	PDF documents	Version	Revision history	Status
Flicker	Flicker	Flicker	Flicker	Flicker	Flicker	Flicker	Flicker
MAExplorer	MAExplorer	MAExplorer	MAExplorer	MAExplorer	MAExplorer	MAExplorer	MAExplorer
ProtPlot	Protplot	TMAP (ProtPlot)	ProtPlot	ProtPlot	ProtPlot	ProtPlot	---
xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx

Summary

- Open2Dprot is a fully open-source n-D proteomics data-mining project for a variety of proteomic expression data sources and is being developed at http://open2dprot.sourceforge.net/
- It has a flexible pipeline-modules project design using XML/RDBMS-caches and portable Java and